DOCUMENT RESUME

ED 090 259 TM 003 533

AUTHOR	Koehler, Roger A.
TITLE	Over Confidence on Probabilistic Tests.
PUB DATE	Apr 74
NOTE	13p.; Paper presented at the American Educational Research Association Annual Meeting (Chicago, Illinois, April 15-19, 1974)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*Confidence Testing; *Guessing (Tests); Measurement Techniques; *Multiple Choice Tests; *Risk; *Scoring Formulas; Testing Problems; Test Interpretation; Test Reliability; Test Validity

ABSTRACT

A potentially valuable measure of overconfidence on probabilistic multiple-choice tests was evaluated. The measure of overconfidence was based on probabilistic responses to nonsense items embedded in a vocabulary test. The test was administered under both confidence response and conventional choice response directions to 208 undergraduate educational psychology students. Measures of vocabulary knowledge based on confidence and choice responses, overconfidence, and risk-taking propensity were obtained. The results indicated that overconfidence was significantly related in a negative direction to probabilistic vocabulary scores. A moderate correlation was found between overconfidence and risk-taking propensity. However, the scatter plot for these measures showed that this relationship may have been spurious. (Author)

18.11

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 090259

Over Confidence on Probabilistic Tests

Roger A. Koehler
University of Nebraska

Paper presented at the Annual Meeting of

the American Educational Research Association

April 1974

Session Number 18.11

TM 003 533

OVER-CONFIDENCE ON PROBABILISTIC TESTS

ROGER A. KOEHLER
University of Nebraska

Probabilistic or confidence testing has been recommended (e.g., Shuford, Albert, & Massengill, 1966) as a more reliable and more valid response procedure for objective examinations than the conventional choice-response method. Through the assignment of subjective probabilities as to the attractiveness of each item alternative, probabilistic testing is designed to remove the guessing factor from objective tests, and to also measure various degrees of partial information. Numerous procedures for obtaining test scores based on confidence assignments have been proposed. However, Shuford et. al. have suggested that probabilistic responses will yield maximal expected scores if and only if a "reproducing" scoring function is utilized to obtain item scores. The use of such scoring functions encourages examinees to be "honest" in their expression of subjective probabilities by yielding a severe penalty when high confidence is assigned to an incorrect alternative.

The literature (e.g., Rippey, 1970; Romberg & Shepler, 1968; Hambleton, Roberts & Traub, 1970; Koehler, 1971) indicates no consistent trends with respect to the improvement of test reliability and/or validity through probabilistic response procedures de Finetti (1965) suggested that the success of probabilistic response procedures is dependent upon examinee understanding of the item scoring function and the expected pay-offs under various degrees of uncertainty. Perhaps the conflicting results of previous studies is partially attributable to a lack of adequate training in confidence response methodology. If, through extensive training, higher reliabilities can be obtained when confidence tests are used

in place of conventional choice tests, an important question remains unanswered: What produces the increased reliabilities? Are they due to the more precise nature of measurement through subjective probability assignment, or could such increases occur as a result of reliably measuring some dimension or trait in addition to the trait the test was designed to measure? If an affirmative answer was given to the latter question, one would have a difficult time arguing in favor of confidence testing procedures.

The purposes of the present study, therefore, were to develop a measure of "over-confidence" on probabilisitc tests, to assess the measurement characteristics of such a measure, and to investigate the relationship of over-confidence on tests to knowledge and to risk-taking propensity. Several authors (e.g., Echternacht, 1972; Stanley & Wang, 1970; Hansen, 1971) have implied that over and/or under confidence expressed by examinees responding to test items through confidence marking procedures can be equated to risk-taking propensity.

METHOD

The experimental instrument for this study was a 40 item multiple-choice vocabulary test. Randomly placed within these 40 items were seven nonsense items, where a nonsense item is defined as an item that has no correct nor incorrect answer. An example of a nonsense item on the test is as follows:

22. Bilious: sad____ double____ greedy_____ bitter_____

Since "Bilious" has no meaning in the English language, the above item has no best (correct) answer and no incorrect answer.

Vocabulary Measures:

Vocabulary scores on the 33 legitimate test items were obtained from both a confidence assignment administration and a conventional choice administration that employed do-not-guess directions. The confidence response directions requested

examinees to assign their percent of confidence in each alternative to the nearest hundredth, making sure their confidence for all alternatives of an item summed to 100 percent. The vocabulary score for item (j), based on the confidence marking directions was obtained by each of the three "reproducing" scoring functions below:

$$S_{1j} = 2P_k - \sum_{i=1}^{m} P_i^2 \qquad \text{(quadratic)}$$

$$S_{2j} = \begin{cases} (2 + \log P_k)/2 & .01 \leq P_k \leq 1 \\ 0 & 0 \leq P_k < .01 \end{cases} \qquad \text{(logarithmic)}$$

$$S_{3j} = P_k / \left(\sum_{i=1}^{m} P_i^2\right)^{\frac{1}{2}} \qquad \text{(spherical)}$$

where P_i is the probability (confidence) assigned to alternative i, P_k is the confidence expressed in the keyed alternative, and m is the number of alternatives per item. Total confidence response vocabulary scores (S_1, S_2, S_3) were calculated by summing the above item scores over all items. Choice responses to the 33 legitimate vocabulary items administered under conventional do-not-guess directions yielded number right scores (L) and "corrected for guessing" scores (G).

Over-Confidence Measures:

The measure of over-confidence was based on confidence responses to the seven nonsense vocabulary items, where the over-confidence for nonsense item (j) was determined by the formula:

$$C_{4j} = \sum_{i=1}^{m} (P_i - 1/m)^2/(1 - 1/m).$$

C_{4j} ranges from a low of zero (equal probability assigned to each alternative) to a high of one (total confidence assigned to a single alternative). The total over-confidence (C_4) expressed by an examinee was the sum of the C_{4j} values on the seven nonsense items. For comparative purposes, and additional measure of confidence

developed by Hansen (1971) and based on probabilistic responses to legitimate items was calculated as:

$$C_3 = C_T - \hat{C}_T$$

In the latter formula,

$$C_T = (1/33) \sum_{j=1}^{33} \left[m/2(m-1) \sum_{i=1}^{m} |1/m - P_{ij}| \right]$$

and $\hat{C}_T$ is the linear estimate of C_T using S_3 as a predictor variable. C_T is a measure of the degree of certainty expressed through confidence responses to legitimate test items. The procedure for determining C_3 was also employed with S_1 and S_2 as predictors and yielded two additional confidence measures, C_1 and C_2 respectively.

Finally, a measure of risk-taking propensity (R) was calculated as the proportion of nonsense items attempted when the vocabulary test described above was administered under conventional do-not-guess directions. This risk measure has been extensively used in research (e.g., Slakter, 1967, 1968a, 1968b, 1969; Slakter & Koehler, 1968) and has yielded high reliabilities for very few nonsense items.

A summary of the total scores derived from the two administrations (confidence response and conventional do-not-guess) of the 40 item vocabulary test are listed below:

1) over-confidence on nonsense items (C_4)
2) quadratic vocabulary score (S_1)
3) logarithmic vocabulary score (S_2)
4) spherical vocabulary score (S_3)
5) number right vocabulary score (L)
6) "corrected for guessing" vocabulary score (G)
7) risk-taking propensity (R)
8) residual confidence-partialling S_1 from C_T(C_1)

9) residual confidence-partialling S_2 from $C_T(C_2)$

10) residual confidence-partialling S_3 from $C_T(C_3)$

Ss for the study were all available students enrolled in an undergraduate educational psychology course; the sample totaled 208 students. Testing sessions for all Ss went as follows:

1. A training booklet was administered to teach Ss how to respond under confidence marking directions. The training booklet was designed specifically to help Ss become familiar with the following:
    a) the confidence response procedure
    b) the logarithmic scoring function (S_1)
    c) the pay-offs for responding in various manners under several degrees of uncertainty (i.e., illustrations pertaining to the severity of the penalties assessed for expressing high confidence in incorrect alternatives were presented).

    Training was provided only for scoring function S_1 in order to test the conjecture that scoring function familiarity and expected pay-offs are necessary for the success of confidence marking methods. Four contrived vocabulary items were placed at the end of the training booklet for the purpose of evaluating the success of the booklet.
2. The vocabulary test was administered through confidence response directions. At the completion of this administration all test booklets were collected.
3. The vocabulary test with a random reordering of items was administered under conventional do-not-guess directions.

RESULTS AND DISCUSSION

An investigation of the responses to the four contrived vocabulary items at

the end of the training booklet provided evidence that Ss understood the confidence marking procedure. For the extremely simple item, Ss assigned 100 percent confidence to the keyed alternative, for the very difficult item, most Ss equally distributed their confidence among the alternatives, and for the other two items, Ss appeared to distribute their confidence in the expected percentages.

Table 1 presents the means, standard deviations, and coefficient alpha reliability estimates for all scores obtained in the study.

Insert Table 1 about here

An inspection of Table 1 indicates that training in the use of the S_1 scoring function did not yield higher reliabilities for that function over the S_2 or S_3 functions. In addition, reliabilities of confidence response scores were generally about the same as those for the L and G conventional response scores (the largest difference occurred between L and S_2 scores, .85 versus .74 respectively). While the reliability for the confidence measure C_4 was not what one might desire, it must be remembered that C_4 is based on only seven items. Using the Spearman-Brown Prophecy formula, a set of 33 nonsense items should yield reliability of .86 for C_4, which is comparable to the reliabilities of the other scores obtained in the study. A reliability of .87 for R, which is also based on the seven nonsense items is consistent with previous research on this risk measure (e.g., Slakter, 1969; Slakter & Cramer, 1969; Slakter & Koehler, 1968).

Note that the mean of the C_4 confidence measure is quite low (i.e., only 0.70 when the maximum possible C_4 is 7.00). Two factors may have contributed to this low C_4 mean. First, the overall test was very difficult (mean of number right scores was only 14.00 of a possible 33 points). This general difficulty may have forced more Ss into a conservative response position. Secondly, the

formula upon which C_4 is calculated is biased toward the low end of the (0-1) interval; i.e., as confidence increases, C_4 increases at a much slower rate.

Table 2 contains the intercorrelations among all scores obtained through both the confidence and the conventional test administrations.

Insert Table 2 about here

As would be expected, the correlations among the vocabulary test scores S_1, S_2, S_3, L, and G were generally high; about the same magnitude as the reliabilities. The correlations between vocabulary scores obtained under confidence response directions (S_1, S_2, and S_3) were significantly ($\alpha < .01$) correlated in a negative direction with both over-confidence (C_4) and risk-taking propensity (R). This latter finding implies that confidence response vocabulary scores tend to be lower for Ss who are overly confident of their responses or who possess a high propensity for taking risks. Since Ss vary with respect to their confidence expression and/or risk-taking behavior, probabilistic testing methods appear to confound knowledge with these two personality traits. Although "corrected for guessing" (G) scores were also significantly ($\alpha < .01$) related to confidence expression, the strength of association ($r^2 = .04$) was somewhat less than that of the $C_4 - S_1$ ($r^2 = .16$), $C_4 - S_2$ ($r^2 = .20$), and $C_4 - S_3$ ($r^2 = .10$) relationships. If a testing method were to be chosen based on the above results, the conventional testing method using number-right (L) scores would appear to be the most valid procedure, since this testing method yields vocabulary scores that are essentially unrelated to over-confidence and risk-taking propensity.

It is interesting to note that confidence measures C_1, C_2, and C_3 correlated positively (significant at the .01 level) with L scores and G scores. In fact, several correlations between these legitimate item confidence measures and conventional vocabulary scores were of the same magnitude as the correlations between

legitimate item confidence measures and the nonsense item confidence measure (C_4). Perhaps the above finding indicates that the linear regression procedure used to obtain C_1, C_2, and C_3 was not totally successful in removing the knowledge dimension from the C_T scores. Using S_1, S_2, or S_3 to partial the knowledge dimension from C_T scores may not be entirely valid. If over-confidence does account for a portion of the variation in confidence response vocabulary scores, both knowledge variation and confidence variation are removed by the linear regression procedure.

With respect to the relationship between risk-taking propensity and over-confidence, the present study indicates a moderate (significant at the .01 level) relationship between C_4 and R, and essentially zero relationship when C_1, C_2, and C_3 are compared to R. The relationship between C_4 and R may be of a spurious nature, since an inspection of the scatter diagram for these variables revealed a rather skewed distribution for C_4 scores. Most C_4 scores ranged between zero and two, while only a very few relatively high risk takers scored greater than 2.5 on the C_4 measure. In addition, the relationship between C_4 and R may be attributed to the fact that these two measures are based on the same <u>few</u> nonsense items. Based on the above relationships, it would appear that over-confidence and risk-taking propensity are not identical traits as previous authors have suggested.

Since the reliability of the C_4 scores was not as high as the reliabilities of the other scores generated in this study (See Table 1), estimates of the correlations between C_4 and the other scores assuming all measures to be perfectly reliable were calculated and are presented in row one of Table 2. In most cases, these estimates tend to support the conclusions made previously.

The results presented above are subject to the limitations inherent in this study. The most serious limitation involves the experimental instrument (vocabulary test) that was used to assess knowledge and confidence. Since a

vocabulary test bore little relationship to the objectives of the educational psychology course from which Ss were obtained, there may have been minimal incentive for Ss to be completely honest in their expressions of confidence. Therefore, further research regarding the problem described in this study should be performed using grade dependent course examinations.

In summary, the present study describes a potentially valuable disguised measure of over-confidence on objective examinations. This measure, which indicates the degree of confidence a subject possesses over and above that which is due to subject matter knowledge (vocabulary knowledge), was significantly related to probabilistically derived test scores and less highly related to number right conventional test scores. It would appear, therefore, that confidence responding methods produce variability in scores that cannot be attributed to knowledge of subject matter (in this study, vocabulary). If these findings could be generalized to all types of objective tests administered under confidence response directions, one could not recommend such response methods as reasonable alternatives to the conventional rights-only procedure.

In addition, the present study indicates that over-confidence in one's responses to vocabulary test items is not identical to one's propensity to take risks on such test items. The measure of over-confidence described in this study was only moderately related to a measure of risk-taking propensity, and this relationship may have been of a spurious nature.

Further research is necessary to investigate possible relationships between the disguised confidence measure described here and various personality traits of examinees.

REFERENCES

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Matematical and Statistical Psychology, 1965, 18, 87-123.

Echternacht, G. J. The use of confidence testing in objective tests. Review of Educational Research, 1972, 42, 217-236.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. A c mparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. Journal of Educational Measurement, 1970, 7, 75-82.

Hansen, R. The influence of variables other than knowledge on probabilistic tests. Journal of Educational Measurement, 1971, 8, 9-14.

Koehler, R. A. A comparison of the validities of conventional choice testing and various confidence marking procedures. Journal of Educational Measurement, 1971, 8, 297-303.

Rippey, R. M. A comparison of five different scoring functions for confidence tests. Journal of Educational Measurement, 1970, 7, 165-170.

Romberg, T. A., & Shepler, J. L. An experiment involving a probability measurement procedure. Paper presented at the meeting of the American Educational Research Association, Chicago, February, 1968.

Shuford, E. H., Albert, A., & Massengill, H.E. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

Slakter, M. J. Risk taking on objective examinations. American Educational Research Journal, 1967, 4, 31-43.

Slakter, M. J. The effect of guessing strategy on objective test scores. Journal of Educational Measurement, 1968, 5, 217-221. (a)

Slakter, M. J. The penalty for not guessing. Journal of Educational Measurement, 1968, 5, 141-144. (b)

Slakter, M.J. Generality of risk taking on objective examinations. Educational and Psychological Measurement, 1969, 29, 115-128.

Slakter, M. J., & Cramer, S. H. Risk taking and vocational or curriculum choice. Vocational Guidance Quarterly, 1969, 18, 127-132.

Slakter, M. J., & Koehler, R. A. A new measure of risk taking on objective examinations. California Journal of Educational Research, 1968, 19, 132-137.

Wang, M. W. & Stanley, J. C. Differential Weighting: a review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.

Table 1

MEANS, STANDARD DEVIATIONS, AND RELIABILITIES OF ALL SCORES
N = 208

SCORE	MEAN	STANDARD DEVIATION	RELIABILITY
Confidence (C_4)	0.70	0.75	0.57
Quadratic (S_1)	10.90	6.14	0.80
Logarithmic (S_2)	23.04	2.97	0.74
Spherical (S_3)	19.35	3.78	0.82
No. Right (L)	14.00	6.18	0.85
Corrected (G)	10.65	6.95	0.82
Risk (R)	0.39	0.36	0.87
Residuals for $S_1(C_1)$	0.00	0.16	0.82*
Residuals for $S_2(C_2)$	0.00	0.17	0.86*
Residuals for $S_3(C_3)$	0.00	0.14	0.74*

*reliabilities of linear combinations: $C_T - \hat{C}_T$

Table 2

CORRELATIONS AMONG VARIOUS SCORES
N = 208

SCORES	C_4	S_1	S_2	S_3	L	G	R	C_1	C_2	C_3
Confidence (C_4)	---	(-.59)*	(-.69)	(-.45)	(-.22)	(-.31)	(.46)	(.66)	(.54)	(.77)
Quadratic (S_1)	-.40	---								
Logarithmic (S_2)	-.45	.97	---							
Spherical (S_3)	-.31	.98	.94	---						
No. Right (L)	-.15	.79	.71	.83	---					
Corrected (G)	-.21	.88	.80	.91	.96	---				
Risk (R)	.32	-.32	-.32	-.32	.06	-.17	---			
Residuals for $S_1(C_1)$	.45	---	---	---	.36	.31	.07	---		
Residuals for $S_2(C_2)$	.38	---	---	---	.48	.44	.02	---	---	
Residuals for $S_3(C_3)$	.50	---	---	---	.24	.17	.12	---	---	---

*Values in parentheses were determined by using the "correction for attenuation," since C_4 scores had lower reliability than the other scores.